Lieff
Cabraser
Heimann&
Bernstein
Attorneys at Law

SUSMAN GODFREY L.L.P.
A REGISTERED LIMITED LIABILITY PARTNERSHIP

CDAS

[redacted]

VIA ECF
Hon. Ona T. Wang
Daniel Patrick Moynihan United States Courthouse
500 Pearl Street
New York, NY 10007-1312

RE: *Authors Guild et al. v. OpenAI, Inc., et al., and Alter et al. v. OpenAI Inc., et al.,* Nos. 1:23-cv-08292-SHS & 1:23-cv-10211-SHS

Dear Judge Wang:

Pursuant to Rule II.b of Your Honor's Individual Practices, Plaintiffs seek an informal discovery conference concerning six OpenAI custodians. This case is about whether OpenAI's reproduction of pirated books to train its large language models is copyright infringement. The six custodians at issue had key roles in [redacted]

The parties have met and conferred several times over Zoom and telephone. While there is agreement on 24 OpenAI custodians, the parties are at impasse as to the six at issue in this motion. *See* Ex. 1 (email exchange); *id.* at 5-7, 25-27 (explanations). Because OpenAI has shown no significant burden to adding these custodians, and because each likely has relevant and noncumulative records, their addition is proportional and warranted.

I. Legal Standard

Discovery disputes regarding the number of custodians are subject to the general principles of relevance and proportionality. *See Fireman's Fund Ins. Co. v. Great American Ins. Co. of New York*, 284 F.R.D. 132, 135 (S.D.N.Y. 2012); *Delta Air Lines v. Lightstone Grp., LLC*, No. 21-MC-374, 2021 WL 2117247 at *2, (S.D.N.Y. 2021) (Wang, M.J.). Courts routinely require production of additional custodial files where, after a threshold showing of relevance, the party resisting discovery fails to "show[] that the production of all of the requested employees' email communications would be unduly burdensome or that a search of their files would not potentially yield relevant information." *See Capitol Recs., Inc. v. MP3tunes, LLC*, 261 F.R.D. 44, 50 (S.D.N.Y. 2009); *Delta*, 2021 WL 2117247 at *3 (granting motion to compel where party failed to "specify[] the nature or size of the burden . . .").[1] Courts may consider whether the requested

[1] In adjudicating a similar dispute in the *In re Chat GPT Litig.*, No. 23-cv-3223, matter the Northern District of California held that plaintiffs were entitled to 24 custodians and allowed the Plaintiffs additional custodians "to the extent that Plaintiffs can identify any number of additional custodians . . . that Plaintiffs can persuasively argue to possess non-cumulative, unique documents or information." Dkt. 166 at 2.

custodians are likely to have nonduplicative material, and plaintiffs can make this showing through citations to documents and inferences based off a custodian's roles / responsibilities. *See United States v. United Techs. Corp.*, 16-CV-01730, 2020 WL 7339916, at *9–10 (D. Conn. Dec. 14, 2020) (granting motion where custodians received relevant emails or had relevant responsibilities); *Fort Worth Ret. Fund v. J.P. Morgan Chase & Co.*, 297 F.R.D. 99, 107 (S.D.N.Y. 2013).[2]

II. Plaintiffs Have Justified the Inclusion of These Six Additional Custodians.

(1) Shantanu Jain: *OpenAI Technical Staff Member* [redacted]. Jain was one a handful OpenAI employees primarily responsible for [redacted] OpenAI's internal documents show that Jain investigated [redacted]. *See, e.g.*, Ex. 2 at 5-9 ([redacted] Jain was heavily involved in [redacted], *see id.* at 12-15, and [redacted], *see id.* at 18 (discussing with OpenAI's President [redacted]). Jain's custodial file will show [redacted]

(2) Ilya Sutskever: *OpenAI Co-Founder, Former Chief Scientist, and Lead OpenAI Safety Critic*. Sutskever's custodial file is likely to contain unique, relevant documents for at least two reasons. First, Sutskever directed [redacted] *See* Ex. 3 at 1 [redacted] *id.* at 2 ([redacted]); *id.* at 3 [redacted]

Second, Sutskever's custodial file is likely to contain unique and relevant information about whether the AI systems at issue are beneficial or detrimental to the public interest. OpenAI itself recently argued in the *New York Times* action that whether "the technology OpenAI introduced to the world" serves "the public interest generally" is relevant to fair use issues. *NYT* Dkt. 236 at 1. And Sutskever is uniquely situated with respect to this issue. He led OpenAI's "alignment" and "superalignment" efforts, which focused on "managing the[] risks" that OpenAI's AI systems "could lead to the disempowerment of humanity," as well as the problem of ChatGPT's regurgitation of copyrighted material. https://openai.com/index/introducing-superalignment/; https://openai.com/index/introducing-the-model-spec/ ("[a]s a continuation of our work on collective alignment and model safety . . ."); https://tinyurl.com/mvf5xt54 (listing "respect[ing] creators and their rights" as a key "rule" in the "model spec"). Sutskever was so gravely concerned

[2] The mere fact that a custodian's file may be partially duplicative of documents already produced is not grounds for denying their inclusion. *See, e.g., Mount Hawley Insurance Co. v. Felman Producution, Inc.*, 269 F.R.D. 609, 620 (S.D.W.V.2010) (granting motion to compel additional custodians even though it was "highly likely that the [requested custodians] will produce ... duplicates of previously produced materials," because "it is reasonable to believe that they will have additional, highly relevant materials . . .").

about the safety of OpenAI models and its rapid commercialization that he orchestrated the (short-lived) firing of OpenAI's CEO Sam Altman over these issues in November 2023. *See* https://venturebeat.com/ai/openais-leadership-coup-could-slam-brakes-on-growth-in-favor-of-ai-safety/. As OpenAI's safety lead during his tenure, and the internal whistleblower over safety issues at OpenAI, Sutskever will no doubt have unique, relevant documents about whether the AI systems at issue will "serve the public interest generally." *NYT* Dkt. 236 at 1.

(3) Jong Wook Kim: *Member of the OpenAI Technical Staff* [redacted]. Kim was specially involved in identifying [redacted] *See* Ex. 4 at 1([redacted]). In particular, Kim played the unique role [redacted] *See id.* at 2 ([redacted] This appears to be—[redacted]

(4) Cullen O'Keefe: *Former Research Scientist* [redacted] O'Keefe served as [redacted]. He authored [redacted] Ex. 5 at 1-17; *see, e.g., id.* at nn. 4, 7-9, 16-20 ([redacted] These [redacted]. *E.g. id.* at 6, 8, 13.

(5) Qiming Yuan: *OpenAI's "Pretraining Data Lead" and "Dataset Sourcing and Processing Lead" for GPT-4o and GPT-4 Respectively*. Mr. Yuan was involved [redacted] *See* Ex. 6 at 1, 5, 8-9.

(6) Andrew Mayne: *Writer and Former OpenAI "Science Communicator.*" Mayne is a published author and former OpenAI employee whose writings have uniquely focused—[redacted]—on the importance of books as training data and LLMs' abilities to write books (*i.e.*, the types of works at issue here). Just this week, Mr. Mayne tweeted, "As someone who is a novelist . . . and worked at OpenAI on capabilities discovery, I'm convinced that AI will be able to write good novels soon." *See* https://x.com/AndrewMayne/status/1830685142781993100. He has also confirmed the importance of books as training data, writing that "training on entire short stories or novels," is necessary, for example, for "a text transformer to understand plot structure." https://tinyurl.com/ypfh3rak; https://tinyurl.com/2687m7cp. Mr. Mayne's communications at OpenAI about books and AI's ability to write books are relevant to whether OpenAI's reproductions of [redacted] is nontransformative (fair use factor 1) and the extent to which OpenAI has usurped writers' creations to create a machine that will replace them (fair use factor 4). *See Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 918-931 (2d Cir. 1994).

Respectfully submitted,

LIEFF CABRASER HEIMANN
& BERNSTEINS LLP

/s/ Rachel Geman

Rachel Geman

SUSMAN GODFREY LLP

/s/ *Rohit D. Nath*

Rohit D. Nath

COWAN DEBAETS
ABRAHAMS & SHEPPARD
LLP

/s/ Scott J. Sholder

Scott J. Sholder

cc: All Counsel of Record (via ECF)